University of California, Berkeley

U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2008 *Paper* 237

Confidence Intervals for the Population Mean Tailored to Small Sample Sizes, with Applications to Survey Sampling

Michael Rosenblum* Mark J. van der Laan†

*Center for AIDS Prevention Studies, Department of Medicine, University of California, San Francisco, mrosenbl@jhsph.edu

†University of California - Berkeley, laan@berkeley.edu

Confidence Intervals for the Population Mean Tailored to Small Sample Sizes, with Applications to Survey Sampling

Michael Rosenblum and Mark J. van der Laan

Abstract

The validity of standard confidence intervals constructed in survey sampling is based on the central limit theorem. For small sample sizes, the central limit theorem may give a poor approximation, resulting in confidence intervals that are misleading. We discuss this issue and propose methods for constructing confidence intervals for the population mean tailored to small sample sizes.

We present a simple approach for constructing confidence intervals for the population mean based on tail bounds for the sample mean that are correct for all sample sizes. Bernstein's inequality provides one such tail bound. The resulting confidence intervals have guaranteed coverage probability under much weaker assumptions than are required for standard methods. A drawback of this approach, as we show, is that these confidence intervals are often quite wide. In response to this, we present a method for constructing much narrower confidence intervals, which are better suited for practical applications, and that are still more robust than con?dence intervals based on standard methods, when dealing with small sample sizes. We show how to extend our approaches to much more general estimation problems than estimating the sample mean. We describe how these methods can be used to obtain more reliable confidence intervals in survey sampling. As a concrete example, we construct confidence intervals using our methods for the number of violent deaths between March 2003 and July 2006 in Iraq, based on data from the study ''Mortality after the 2003 invasion of Iraq: A cross-sectional cluster sample survey", by Burnham et al. (2006).

1 Introduction

Our paper is motivated by applications in survey sampling where cluster sampling is used, and the number of clusters is relatively small. In this case, standard, normal-based confidence intervals for the population mean may not provide the wished 95% coverage. The goal of this paper is to provide alternative methods for constructing confidence intervals that have improved coverage probability for small sample sizes. We present two sets of alternative methods. Our first set of methods is based on tail bound inequalities, such as Bernstein's inequality, that can be used to construct confidence intervals with correct coverage under much weaker assumptions than required by normal-based methods. Our second set of methods, based on a numerical optimization approach, gives confidence intervals with correct coverage at all sample sizes for certain parametric models (such as the negative binomial model).

All of the methods we propose for constructing confidence intervals are based on so-called "exact" methods. As described in (Blyth and Still, 1983), an "exact confidence interval" is such that the coverage probability is greater than or equal to the nominal level (e.g. 0.95), for all sample sizes and for all possible data generating distributions satisfying the assumptions of the method. Exact methods for constructing confidence intervals, typically based on inverting hypothesis testing procedures, are known for many parametric models (Clopper and Pearson, 1934; Sterne, 1954; Crow and Gardner, 1959; Casella and Berger, 1990). We give exact methods under nonparametric models, based on tail bound inequalities such as Bernstein's inequality. We also give exact methods under parametric models, such as the negative binomial distribution, along with algorithms for computing confidence intervals under such models.

We extend our methods to allow construction of confidence intervals for a rich class of parameters, including, for example, coefficients for linear regression models, for logistic regression models, and for Cox proportional hazards models. This is done by applying our methods to the empirical mean of an estimator's influence curve. This represents a useful generic approach for construction of more conservative and thereby more reliable confidence intervals in many different applications.

At the end of the paper, we lay out a template for future research, aimed at development of new methods tailored to statistical inference in situations with small sample sizes. In particular, we describe how any improvement in tail bounds of sums of independent random variables immediately translates into improved methods for constructing confidence intervals for small sample

sizes.

How small a sample must be in order for standard methods to perform poorly will depend on how skewed or heavy-tailed the data generating distribution is. Thus, there is no clear threshold for when a sample size is considered "small"; however, for concreteness, we use sample size $n = 50$ (i.e. 50 clusters) in our examples.

The organization of this article is as follows. In the next section, we describe the type of survey sampling application motivating our work. In Section 3, we discuss problems with using standard confidence intervals, such as normal-based or bootstrap-based confidence intervals, for small sample sizes. In Section 4, we present the first of our methods: a construction of confidence intervals based on Bernstein's inequality. In Section 5 we show how to construct confidence intervals based on other tail bound inequalities. In Section 6, we give a method for constructing confidence intervals that is less conservative than our methods based on tail bounds, but that is more robust than standard normal-based and bootstrap-based confidence intervals. In Section 7, we compare the widths of confidence intervals based on the different methods in the paper. In Section 8, we show how to generalize our methods to a rich class of parameters, which include, for example, coefficients of Cox proportional hazards models. In Section 9, we illustrate our methods by constructing confidence intervals for the number of violent deaths in a post invasion period in Iraq based on data used in the article "Mortality after the 2003 invasion of Iraq: A cross-sectional cluster sample survey" (Burnham et al., 2006). Lastly, in Section 10, we describe open problems and a research agenda for improving statistical inference for small samples.

2 Survey Sampling using Cluster Sampling Designs

Consider a cluster sampling design involving sampling n clusters of households and counting the number of events in each cluster. We are interested in designs for which the total number of clusters n is relatively small ($n = 50$ for example). For concreteness, we focus on the type of application that will be discussed in Section 9: estimating the number of deaths due to violence in Iraq during a specified period. We first describe this estimation problem in the case where clusters are chosen randomly within a single area. We then consider a more complex cluster sampling scenario, in which one divides up the sampling area into K sub-areas that are assumed to have different rates of deaths due to violence. In both cases, we reduce the problem of estimating

the total number of deaths due to violence during a specified period to the problem of estimating the (population) mean of a sum of n independent random variables (representing the counts in each cluster). The rest of the paper focuses on robust methods for estimating such a mean of independent random variables, when the number of observations n is relatively small. We end this section with a discussion of design-based vs. superpopulation-based methods in survey sampling, and how our methods are examples of design-based methods.

2.1 Cluster Sampling from a Single Population

Consider the problem of estimating the number of deaths due to violence in a large area. Since it is often hard to count all such deaths, we consider the strategy of taking a random sample of n clusters of households and counting the number of violent deaths in these clusters of households. Denote the number of deaths due to violence in each of the n clusters, by the random variables $X_1, \ldots, X_n$. Assume the clusters are randomly selected in such a way that $X_1, \ldots, X_n$ are independent, identically distributed, and such that for each cluster i, each household has the same probability of being selected in it. Denote the mean of X_1 by μ.

We describe how to map an estimate for the mean count of deaths due to violence in a cluster into an estimate for the total number deaths due to violence in the area. Let N be the total size of the population living in the area, which we assume is (approximately) known. Let D be the (unknown) total number of deaths due to violence that occurred in the area over the specified period. Let C_i be the random variable denoting the ith cluster of individuals (e.g. identified by name and house address or GPS location). Let $|C_i|$ denote the total number of people in the ith cluster. Then it follows by the assumption in the previous paragraph that the total number of deaths due to violence D equals $NEX_1/E|C_1|$. That is, based on an estimate of μ and an estimate of the mean cluster size, we can form an estimate of the number of deaths due to violence in the total area.

A natural estimate of μ is the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. The rest of this paper is devoted to constructing confidence intervals for μ (which we can then extrapolate to a confidence interval for the total number of violent deaths D), that are appropriate when n is relatively small (e.g. $n = 50$).

2.2 Cluster Sampling from Multiple Populations

Here we consider cluster survey designs in which one divides up the sampling area into K sub-areas that are assumed to have different rates of deaths due to violence. This approach is useful when there is a priori knowledge that violent death rates will differ by region.

For each sub-area $k \in \{1, \ldots, K\}$, let $n(k)$ denote the number of clusters sampled in that sub-area. Let X_k denote the random variable representing the number of deaths due to violence in a randomly chosen cluster in the kth sub-area. Let $X_{k1}, \ldots, X_{kn(k)}$ denote i.i.d copies of X_k, for each $k \in \{1, \ldots, K\}$. These represent the number of deaths due to violence counted in each of the $n(k)$ clusters in each sub-area k. Let $\mu(k)$ be the mean of X_k, representing the mean number of deaths due to violence in a randomly chosen cluster in the kth sub-area. Let $w(k)$ be a weight proportional to the population size of the kth area, for each $k \in \{1, \ldots, K\}$, and assume these weights sum to 1.. Let $\mu = \sum_{k=1}^{K} w(k)\mu(k)$. Assuming that the mean cluster size is the same in all sub-areas, and choosing weights $w(k)$ proportional to the population in each sub-area, we show that μ can be mapped into the total count of deaths due to violence for the entire sampling area.

Under similar assumptions as in the previous subsection, and letting C_k denote a randomly chosen cluster in the kth sub-area, we have

$$D_k = N_k \frac{\mu(k)}{E|C_k|},$$

where N_k is the population size of sub-area k, and D_k is the number of violent deaths in this sub-area. Also, denote the total population in all sub-areas combined, $\sum_{k=1}^{K} N_k$, by N. Assume the mean cluster size is the same in all sub-areas, and denote the mean cluster size by $E|C|$. Then, the total number of violent deaths over the total area is given by the formula:

$$D = \sum_{k=1}^{K} D_k = \sum_{k=1}^{K} \frac{N_k}{E|C|}\mu(k).$$

As a consequence, using weights $w(k) = N_k/N$, one has that the weighted average $\mu = \sum_{k=1}^{K} w(k)\mu(k)$ is equal to $D\frac{E|C|}{N}$. Thus, given an estimate and confidence interval for μ, we can extrapolate to an estimate and confidence interval for the desired D by multiplying by $N/E|C|$.

Let $\bar{X} = \sum_{k=1}^{K} w(k)\bar{X}_k$ be the empirical estimate of μ, where $\bar{X}_k = \frac{1}{n(k)} \sum_{i=1}^{n(k)} X_{ki}$ is the sample mean for sub-area k, for each $k \in \{1, \ldots, K\}$.

The variance of this empirical estimate $\bar{X}$ is

$$VAR(\bar{X}) = \sum_{k=1}^{K} (w(k)^2/n(k)) VAR(X_k). \quad (1)$$

We now consider a special property of the empirical estimate $\bar{X}$ of μ, in the case in which clusters are allocated proportional to population size, as was done in the cluster sampling survey of mortality in Iraq of Burnham et al. (2006) that we consider in Section 9.[1]

Consider the special case in which the number of clusters in each sub-area is proportional to the population size of the sub-area. This proportionality means that the weights $w(k) = N_k/N = n(k)/n$. In this case, we have that our empirical estimator $\bar{X}$ for μ can be written in the following convenient form:

$$\bar{X} = \sum_{k=1}^{K} w(k)\bar{X}_k = \sum_{k=1}^{K} \frac{n(k)}{n} \frac{1}{n(k)} \sum_{i=1}^{n(k)} X_{ki} = \frac{1}{n} \sum_{k,i} X_{ki}. \quad (2)$$

Thus, the variance of our estimator $\bar{X}$ equals the variance of $\frac{1}{n}\sum_{k,i} X_{ki}$, which is $\frac{1}{n^2}\sum_{k,i} \mathrm{VAR}(X_{ki})$. Denote the quantity $\frac{1}{n-1}\sum_{i,k}(X_{ki} - \bar{X})^2$ by s^2, which represents what the standard unbiased estimate for the variance would be if we were to treat $\{X_{ik}\}$ as i.i.d. It is straightforward to prove that Es^2/n is greater than or equal to $\frac{1}{n^2}\sum_{i,k} \mathrm{VAR}(X_{ik})$, with equality only when the mean of X_k is the same for all sub-areas k. Thus, s^2/n is generally an overestimate for the variance of $\bar{X}$. We will use this fact as well as the fact that (2) holds when clusters are allocated proportional to population size, in applying our methods to the data example of Section 9.

2.3 Design-Based vs. Superpopulation-Based Inference in Survey Sampling

The methods for constructing confidence intervals in this paper are examples of design-based inference in survey sampling. This is in contrast to superpopulation-based inference. We briefly describe the difference between these two survey sampling approaches and how our methods fit in. More detailed comparisons of these two approaches can be found in (Chaudhuri, 2005).

[1]The design in (Burnham et al., 2006) involves several levels of systematic and random allocation of clusters proportional to population size. Their design is discussed in Section 9.

In design-based inference, one assumes a fixed population from which a sample is randomly selected according to a particular design. The methods described in this paper all assume such a setup, where the design is either i.i.d. draws from a single population (as in Section 2.1) or from multiple populations (as in Section 2.2). In the data example of Section 9, we assume that in each Governorate (sub-area), clusters are drawn i.i.d. from the population of all possible clusters that could be chosen in that Governorate.[2] We assume that for each possible cluster that could be chosen, there is a fixed count representing the number of deaths due to violence in that cluster. As in design-based inference in general, we assume the only source of randomness is due to the manner in which clusters are selected.

In contrast to design-based inference, superpopulation-based inference (also called model-based inference) assumes that the population being sampled is itself a random draw from a distribution called the "superpopulation." One possible superpopulation-based inference approach to the data example of Section 9 would be to assume a statistical model for the mortality rates in each Governorate for different periods of time. Such a model may assume that for different periods within the study time frame, these mortality rates within a Governorate are correlated, and the model would involve assumptions about such correlations. In contrast, our design-based approach does not assume the mortality rates are random–only that the choice of clusters within Governorates is random. We therefore make no assumptions on, for example, correlations over time of mortality rates within clusters. In fact, we use no information in our analysis beyond the sample mean, sample variance, and maximum value of the aggregate cluster counts; we thus do not exploit or model any within-cluster characteristics.

3 Problems with Using Standard Confidence Intervals with Small Samples

In the previous section, we reduced the problem of estimating a total count, such as the total number of deaths due to violence in Iraq over a period of time, to the problem of estimating the mean of a sum of a relatively small number of independent random variables. Confidence intervals for such a mean based

[2]This is an oversimplification, since the actual study design involved two levels of allocation of clusters; clusters were first allocated systematically to Governorates, and then allocated randomly proportional to population size to administrative units within Governorates. We ignore this second level of allocation and make the simplifying assumption that within each Governorate, clusters are allocated i.i.d. within that Governorate.

on the normal distribution, Student's t-distribution, Poisson distribution, or bootstrap, may have poor performance in such small samples. In this section we give several examples of how bad this performance can be. In particular, we give examples of distributions for which 0.95 confidence intervals based on these methods contain the population mean with probability much less than 0.95. Before giving these examples, we briefly review the theoretical justification for confidence intervals based on the normal distribution and Student's t-distribution. This exposes the assumptions underlying these methods, which may not hold in many cases.

3.1 Assumptions Underlying Standard Methods

Suppose one samples n independent and identically distributed (i.i.d.) draws $X_1, \ldots, X_n$ of a random variable X with mean μ and variance σ^2, and suppose that one wishes to construct a 0.95 confidence interval for μ.

A natural estimate of the mean μ of X is the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. It is common practice to estimate the standard error of $\bar{X}$ with $s/\sqrt{n}$, where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is an unbiased estimate of the variance. Then one reports as 0.95-confidence interval

$$\bar{X} \pm z\frac{s}{\sqrt{n}},$$

where z is chosen to be 1.96 or to be the 0.975-quantile of Student's t-distribution with $n-1$ degrees of freedom (which we denote by $t_{n-1}(0.95)$). The claimed 0.95 coverage of this confidence interval corresponds with stating

$$0.95 \leq P\left(\bar{X} - zs/\sqrt{n} \leq \mu \leq \bar{X} + zs/\sqrt{n}\right). \tag{3}$$

If X is normally distributed and z is set to be $t_{n-1}(0.95)$, then this is a true statement. However, when X is not normally distributed, such a confidence interval is only an approximation, and may be a poor approximation for small sample sizes.

For (3) to be a good approximation, one generally needs that both:

- Assumption 1: The sample mean of the data generating distribution is approximately normally distributed, and
- Assumption 2: The standard error of the point estimate is accurately estimated.

Neither of these assumptions is guaranteed to be true for small samples; in Examples 1 and 2 below, we give situations where Assumption 1 and Assumption 2 are violated, and where confidence intervals based on (3) contain the true mean with probability less than 0.95.

We point out that standard methods for constructing confidence intervals may fail, in particular, when one is confronted with a highly skewed, heavy tailed probability distribution and one is only able to observe a relatively small sample from this distribution (Wilcox, 2005). In this case, not only will normal-based methods give poor coverage, but also parametric bootstrap-based methods may perform poorly. Methods based on the nonparametric bootstrap can also be very limited in their performance due to the fact that such a small sample will not be representative: that is, the empirical distribution putting probability $1/n$ on each of the n counts $x_1, \ldots, x_n$ will be a very poor approximation of the true probability distribution P. Example 1 below gives a data generating distribution where the nonparametric bootstrap performs poorly, and Example 3 gives a data generating distribution where the parametric bootstrap performs poorly.

3.2 Examples When Standard Confidence Intervals Have Poor Coverage

We now present three examples in which standard confidence intervals have poor performance. In all of the examples, the data generating distribution is highly skewed, though in some cases this will often not be detectable from looking at the data. The first two examples illustrate how each of Assumptions 1 and 2 above can be violated, leading to poor coverage; Example 3 shows how a model-based method can fail when the model is misspecified. In examples 1 and 3, we show how to construct skewed data generating distributions for which the coverage of standard methods will be arbitrarily low.

Example 1: Poor Coverage Due to Underestimation of Variance We give an example of a random variable with continuous distribution for which nominal 0.95 confidence intervals based on the normal distribution, on Student's t-distribution, or on the nonparametric bootstrap contain the true mean with probability much less than 0.95. To construct this random variable, first let δ be a number in $(0, 1)$. Let A take value 0 with probability $1 - \delta$ and take value 1 with probability δ. Let Y be a zero-mean, normal random variable with variance τ^2. Assume A and Y are independent. Let $X = A + Y$. Then the mean of X is δ. Assume we observe n i.i.d. copies of X. Then for any sample size n, we can choose δ and τ so that the 0.95 confidence intervals

based on the normal distribution, Student's t-distribution, or nonparametric bootstrap will contain the mean of X with arbitrarily low probability. We prove this in Theorem 3 in Appendix 1.

To be more concrete, consider sample size $n = 50$. Let $\delta = 0.01$ and $\tau = 0.032$. We show in Appendix 1 that standard 0.95 confidence intervals based on the normal distribution, on the Student's t-distribution, or on the nonparametric bootstrap actually contain the mean of X with probability less than 0.67. In other words, the confidence interval that ought to contain the actual mean with probability at least 0.95, in fact only contains it with probability less than 0.67. The distribution giving rise to this poor performance is highly skewed, though in many data sets of size 50 drawn from this distribution this will not be apparent or detectable by any diagnostic test; the reason is that with probability 0.605, all 50 copies of A equal 0, and so the data are $Y_1, \ldots, Y_n$, which are normally distributed. In this example, the main reason that standard confidence intervals fail to contain the true mean with probability 0.95 is that the estimated variance will quite often severely underestimate the true variance. That is, Assumption 2 above is severely violated.

As another case, consider sample size $n = 50$ and let $\delta = 0.12$ and $\tau = 0.01$. Consider by what multiplicative factor the normal-based interval width would have to be uniformly inflated in order to guarantee 0.95 coverage. A simulation in R shows that one would need to increase the confidence interval width by a multiplicative factor of more than 1.3 in order to get 0.95 coverage for this data generating distribution. Here the problem is not that all 50 copies of A are 0 with probability more than 0.05 as in the previous paragraph–in fact, the probability of all 50 copies of A being 0 is only 0.0016.

Example 2: Poor Coverage Due to Sample Mean Not Approximately Normal

By choosing δ and τ differently in the data generating distribution defined in the previous example, we can cause a severe violation of Assumption 1 that leads to poor coverage probability. We assume the variance of X, denoted by σ^2, is known. We show using numerical computation in Appendix 1 that for any sample size $n : 1 < n < 1000000$, we can choose δ and τ such that the normal-based confidence interval

$$\left(\bar{X} - 1.96\sigma/\sqrt{n}, \bar{X} + 1.96\sigma/\sqrt{n}\right) \tag{4}$$

contains the mean of X with probability less than 0.84. Furthermore, we consider how much the 1.96 in the previous display would have to be increased in order for this confidence interval to have coverage probability 0.95 for a set of data generating distributions we construct; we show that for sample size $n = 50$, the 1.96 in the previous display would have to be increased to 4.18

in order for (4) to have coverage probability 0.95 for the set of distributions defined in the previous example. This would entail more than doubling the size of the standard confidence interval.

For $n = 50$, if we choose $\delta = 0.0035$ and $\tau = 0.0001$, then the probability that (4) contains the true mean is approximately 0.839 (computed by simulation in R described in Appendix 1). In this example, when the confidence interval fails to contain the true mean, it is most often because the confidence interval is to the right of the true mean, so that the entire confidence interval is an overestimate.

This example shows that the assumption that the distribution of the sample mean is sufficiently close to normal may be violated, and may lead to poorly performing confidence intervals, even when the variance is known. Here this is due to the distribution of the sample mean for the binomial distribution (that is, for the distribution of $\sum A_i$) being non-normal when the probability of $A_i = 1$ is small relative to the sample size. We next turn to an example showing that model-based methods using the parametric bootstrap can perform poorly.

Example 3: Poor Coverage from the Parametric Bootstrap
We show in this example how model-based methods can give poor coverage when the model is misspecified. More precisely, we construct a random variable taking values in $0, 1, 2, \ldots$, and show that confidence intervals based on the Poisson distribution using the parametric bootstrap fail to contain the true mean with high probability. The main reason for this failure, in our example, is that the parametric model is misspecified, which will generally be the case in practice.

Let Y be a Poisson random variable with mean 1. Let t be a positive integer. Let A be independent of Y and take value 0 with probability $1 - \delta$ and take value t with probability δ. Let $X = A + Y$. Assume we observe n i.i.d. copies of X.

Then for any sample size n, if δ is sufficiently small and t sufficiently large, the 0.95 confidence intervals based on the Poisson distribution using the parametric bootstrap will fail to contain the mean of X with high probability. We prove this in Theorem 4 in Appendix 1.

Consider sample size $n = 50$. Then for $\delta = 0.01$ and $t = 50$ the probability that 0.95 confidence intervals based on the Poisson distribution using the parametric bootstrap contain the actual mean of X is less than 0.50.

An important question is whether, for the distributions encountered in practice, confidence intervals based on the central limit theorem or on the bootstrap will perform well or poorly given small sample sizes. In the examples above in which these methods performed poorly, the data generating distributions were highly skewed. Wilcox (2005) considers published papers

that attempt to characterize the skewness of distributions that researchers are likely to encounter in practice. Wilcox (pg. 110) remarks that "The most striking feature of these studies is the extent to which they differ. For example, some papers suggest that distributions are never extremely skewed, while others indicate the exact opposite." Even if our examples above are more skewed than would be seen in practice, they do highlight the issue that normal-based and bootstrap-based confidence intervals have the potential to fail severely. Furthermore, a key feature of these examples is that when dealing with small sample sizes, it is often not possible to determine from looking at the data whether the underlying distribution is approximately normal or is highly skewed. Thus, it cannot in general be determined from the data whether standard methods will perform well or not for the application at hand.

4 Construction of Confidence Intervals based on Bernstein's Inequality

The above examples show how normal-based and bootstrap-based confidence intervals can have poor coverage for small sample sizes. We present our first method for generating confidence intervals that is guaranteed to have coverage probability at least 0.95, under milder assumptions than are required for normal-based or bootstrap-based methods.

We rely on Bernstein's inequality, which bounds the tails of the distribution of the sample mean of n independent, bounded random variables $X_1, \ldots, X_n$. The tail bound is in terms of the sample size n, the maximum deviation W the random variable can have from its mean, and an upper bound v on the sum of the variances of the random variables. Bernstein's inequality is the following (see (Dudley, 1999) for a proof): for all $x > 0$,

$$P\left(\frac{1}{n}|\sum_{i=1}^{n}(X_i - EX_i)| > x/\sqrt{n}\right) \leq 2\exp\left(-\frac{nx^2}{2(v + W\sqrt{n}x/3)}\right), \quad (5)$$

where W is a constant for which, for all $i \leq n$, $P\left(|X_i - EX_i| \leq W\right) = 1$, and $v \geq \sum_{i=1}^{n} \mathrm{VAR}(X_i)$

If W and v are known, then we can construct a confidence interval for $\mu := \frac{1}{n}\sum_{i=1}^{n} EX_i$ by letting x^* be the unique positive value of x such that the right hand side of (5) equals 0.05. Then we have by (5) that

$$0.95 \leq P\left(\bar{X} - x^*/\sqrt{n} \leq \mu \leq \bar{X} + x^*/\sqrt{n}\right) \quad (6)$$

We give a formula for calculating x^* in the following lemma:

Lemma 1 *For any significance level $\alpha \in (0,1)$, and any $W > 0$, $v > 0$, and $n > 0$, we define $q(1-\alpha, W, v)$ to be the positive solution q of*

$$2\exp\left(-\frac{nq^2}{2(v + W\sqrt{n}q/3)}\right) = \alpha,$$

which is given by

$$q(1-\alpha, W, v) \equiv \frac{\ln(2/\alpha)}{n}\left(\frac{W\sqrt{n}}{3} + \sqrt{\frac{W^2 n}{9} + \frac{2vn}{\ln(2/\alpha)}}\right). \tag{7}$$

Although the function q depends on n, we omit this variable in what follows to simplify the notation. Note that x^* defined above equals $q(0.95, W, v)$.

Bernstein's inequality requires that the random variables $X_1, \ldots, X_n$ be independent, but does not require they be identically distributed; this will come in handy in dealing with cluster surveys where clusters are allocated to different sub-areas, as discussed in Section 2.

We now consider the special case in which $X_1, \ldots, X_n$ are i.i.d. The following theorem gives Bernstein-based confidence intervals based on the sum of n i.i.d., bounded random variables.

Theorem 1 *Let $X_1, \ldots, X_n$ be i.i.d. copies of a random variable X with mean μ. Let W be such that $P(|X - \mu| \le W) = 1$. Let σ^{*2} be a number satisfying $\sigma^{*2} \ge VAR(X)$. Then, for the function q defined in (7),*

$$P\left(\bar{X} - \frac{q(1-\alpha, W, n\sigma^{*2})}{\sqrt{n}} \le \mu \le \bar{X} + \frac{q(1-\alpha, W, n\sigma^{*2})}{\sqrt{n}}\right) \ge 1 - \alpha.$$

In particular,

$$\left(\bar{X} - q(0.95, W, n\sigma^{*2})/\sqrt{n}, \bar{X} + q(0.95, W, n\sigma^{*2})/\sqrt{n}\right)$$

is a 0.95 confidence interval for μ.

The theorem follows directly from Bernstein's inequality (5). It gives a recipe for creating a 0.95 confidence interval for μ, if a bound W on $|X - \mu|$ and a bound σ^{*2} on the variance of X are both known. However, in practice these will generally not be known, and so must be approximated. The accuracy of these approximations can affect the coverage probability of the resulting confidence intervals. We now discuss methods for such approximations.

The bound W should, at minimum, be set to a value larger than the maximum absolute deviation of all values in the data set from the sample mean.

That is, W should be set to be at least as large as $\max_i |X_i - \bar{X}|$. However, this will tend to underestimate the maximum absolute deviation from the mean. If other information is available to help determine the maximum absolute deviation from the mean (such as past studies or scientific knowledge of the processes generating the data), this can be used as well to help determine an appropriate bound W. As an oversimplified example, if the outcome is systolic blood pressure, it is known that this must lie within a certain interval, say $[a, b]$; then W could be taken as $b - a$. We give an example of how information from past studies is used to select a suitable W in Section 9.

We now turn to approximating an upper bound σ^{*2} on the variance of X. Setting σ^{*2} to be the value of s^2 in Theorem 1 will generally result in more conservative confidence intervals than those based on normal or bootstrap methods; this is shown in Section 7. In practice, this may be sufficient. However, s^2 is not in general an upper bound for σ^2, since s^2 is merely an estimate of σ^2. Generally, s^2 will itself have a large variance, so will be less than σ^2 with non-negligible probability. Therefore, it is advantageous to do a sensitivity analysis in which larger values than s^2 are used for σ^{*2} in Theorem 1. This can be done by estimating the standard error of s^2. Some possible choices for the upper bound σ^{*2} are s^2 plus one or two times this standard error estimate. We describe a method for such an estimate, based on the influence curve of s^2, in Appendix 3, where we also give R-code to compute it. In general, we recommend evaluating the sensitivity of the reported confidence interval with respect to both the choice of W and the choice of σ^{*2}, using the R-code in Appendix 3 for example.

A straightforward generalization of Theorem 1 applies to sums of independent random variables that are not necessarily identically distributed. The only differences are that now $\mu := \frac{1}{n}\sum_{i=1}^{n} EX_i$ and σ^{*2} in the statement of Theorem 1 must instead be an upper bound for $\frac{1}{n}\sum_{i=1}^{n} \mathrm{VAR}(X_i)$.

We can apply the above method for generating confidence intervals to the survey sampling application described in Section 2. In that section, we reduced the problem of estimation and inference for the total number of deaths due to violence in Iraq during a certain time period to the problem of estimation and inference for the mean of a (weighted) sum of independent random variables. These random variables were the total counts in each of the randomly chosen clusters. Recall that we denoted the cluster counts for each cluster in sub-area $k \in \{1, \ldots, K\}$ by $X_{k1}, \ldots, X_{kn(k)}$. For each sub-area k, these counts $X_{k1}, \ldots, X_{kn(k)}$ are assumed to be i.i.d. copies of a random variable X_k. For each sub-area k, we let the weight $w(k)$ be the proportion of the total population that is in sub-area k. The following theorem presents a Bernstein-based 0.95-confidence interval for $\mu = \sum_{k=1}^{K} w(k) EX_k$, which can be directly

extended, as discussed Section 2, to a confidence interval for the total number of deaths due to violence in the entire study area.

Theorem 2 *Consider, for each sub-area* $k \in \{1, \ldots, K\}$*, the sample* $X_{k1}, \ldots, X_{kn(k)}$ *of i.i.d. copies of* X_k*. Let* $\mu(k) = EX_k$*. Let* $\mu = \sum_{k=1}^{K} w(k)\mu(k)$ *and* $\bar{X} = \sum_{k=1}^{K} w(k)\bar{X}_k$*, where* $\bar{X}_k = \frac{1}{n(k)} \sum_{i=1}^{n(k)} X_{ki}$*. Let* $n = \sum_{k=1}^{K} n(k)$*. Let* W *be a positive value such that* $P\left(\frac{nw(k)}{n(k)}|X_k - \mu(k)| \leq W\right) = 1$*. Let* $\sigma^{*2} = n^2 \sum_{k=1}^{K} \frac{w(k)^2}{n(k)} \sigma^{*2}(k)$*, where for all* k*,* $\sigma^{*2}(k) \geq VAR(X_k)$*.*

Then

$$0.95 \leq P\left(\bar{X} - q(0.95, W, \sigma^{*2})/\sqrt{n} \leq \mu \leq \bar{X} + q(0.95, W, \sigma^{*2})/\sqrt{n}\right),$$

so that the corresponding interval is a 0.95 *confidence interval for* μ*.*

The theorem follows directly from Bernstein's inequality (5) and the fact that

$$VAR(\bar{X}) = \sum_{k=1}^{K} (w(k)^2/n(k)) VAR(X_k). \tag{8}$$

Note that the factor $nw(k)/n(k)$ in the requirement on W above equals 1 when the number of clusters in each sub-area is chosen proportional to the population of the sub-area. The above theorem guarantees 0.95 coverage of Bernstein-based confidence intervals regardless of the number of clusters allocated to each sub-area.

We point out that when clusters are allocated proportional to population size, as argued at the end of Section 2, we have that our estimator for μ, $\bar{X} = \sum_{k=1}^{K} w(k)\bar{X}_k = \frac{1}{n} \sum_{k,i} X_{ki}$, and that s^2/n is generally an overestimate of the variance of $\bar{X}$ (where s^2 was defined as the standard unbiased estimator of the variance if we treat $\{X_{k,i}\}$ as i.i.d.). Thus, we can treat the data as if it were i.i.d., and use ns^2 as an approximate upper bound on the variance of the sum $\sum_{i,k} X_{ik}$, to get an approximate Bernstein-based confidence interval. This is especially useful when some sub-areas have few clusters. We will use the approach just outlined in Section 9.

5 Tail Bounds Based on Inequalities of Bennett and Hoeffding

We show how to leverage other inequalities to obtain small sample confidence intervals analogous to those described above based on the Bernstein inequality.

These inequalities bound the tails of sums of independent random variables. We consider two such inequalities, and the assumptions they require:

1. Bennett's inequality, which is an improvement on Bernstein's inequality, requires that a bound on the maximum absolute deviation from the mean and a bound on the variance are known.

2. An inequality of Hoeffding we present only requires that a bound on the maximum absolute value of the data generating distribution is known. We refer to this inequality as "Hoeffding's inequality" in this paper.

The advantage of having these different tail bounds is that one can choose the strongest bound for any given situation. We describe how to do this after we present each of the bounds.

Bennett's inequality (Bennett, 1962, 1963) (see (van der Vaart, 1998) Appendix 6 for an overview) requires the same assumptions required for Bernstein's inequality, that is, X_i are independent random variables, there is a known bound W such that for all i, $P(|X_i - EX_i| \leq W) = 1$, and there is known upper bound v on the variance of $\sum_{i=1}^{n} X_i$. It states that for $\mu = \frac{1}{n}\sum_{i=1}^{n} EX_i$, for all $x > 0$, we have

$$P\left(|\bar{X} - \mu| > x/\sqrt{n}\right) \leq 2\exp\left[-\frac{v}{W^2}\theta\left(\frac{\sqrt{n}xW}{v}\right)\right] \tag{9}$$

where

$$\theta(x) = (1+x)\ln(1+x) - x.$$

Thus,

$$\left(\bar{X} - \frac{x^*}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{x^*}{\sqrt{n}}\right) \tag{10}$$

is a 0.95 confidence interval for μ if x^* is chosen so that the right hand side of (9) is at most 0.05. x^* can be easily computed using a root-finding algorithm (such as **uniroot** in R). Bennett's inequality gives confidence intervals with narrower widths than Bernstein's inequality, under the same assumptions as we stated for Bernstein's inequality, and so should always be used instead of Bernstein's inequality.

Hoeffding's inequality (Hoeffding, 1963) (see Appendix 6 of (van der Vaart, 1998) for an overview) implies for $X_1, \ldots, X_n$ independent random variables, and W such that for all i, $|X_i| \leq W$ with probability 1, and $\mu = \frac{1}{n}\sum EX_i$, we have for all $x > 0$ that

$$P\left(|\bar{X} - \mu| > x/\sqrt{n}\right) \leq 2\exp(-x^2/(2W^2)). \tag{11}$$

Thus, it follows that

$$P\left(\bar{X} - x/\sqrt{n} \leq \mu \leq \bar{X} + x/\sqrt{n}\right) \geq 1 - 2\exp(-x^2/(2W^2)), \qquad (12)$$

A 0.95 confidence interval is obtained setting x in the previous display equal to $2.72W$ (so that the right hand side equals 0.95). This is an improvement, though a modest one, on what one gets when using Bernstein's or Bennett's inequality with W^2 as a bound on the variance of each X_i. In Section 7, we compare the widths of confidence intervals generated based on the above methods.

One can take full advantage of the above inequalities by using all of them to compute confidence intervals, and then taking the narrowest such interval. Note that all the resulting confidence intervals are centered at the sample mean. No multiple testing adjustment is needed, since taking the narrowest interval corresponds to using the strongest tail bound for the particular sample size and known characteristics of the distribution (such as upper bounds on the variance, maximum deviation from the mean, and maximum absolute value).

6 Small Sample Exact Confidence Intervals Based on Parametric Models

In this section, we present our second set of methods for constructing confidence intervals. They are less conservative than the methods in Sections 4 and 5, but are still more robust than standard normal-based and bootstrap-based confidence intervals. The primary motivation for such a method is that, as we investigate in Section 7, the widths of confidence intervals based on the Bernstein inequality and other tail bound inequalities can be much larger than the widths corresponding to normal-based methods.

Let $\mathcal{M}$ denote a statistical model, such as a negative binomial model. The method of this section involves constructing confidence intervals for the population mean that have correct coverage probability whenever the data is generated according to a distribution in the model $\mathcal{M}$. The model $\mathcal{M}$ should be chosen based on subject-matter knowledge for the application at hand. In selecting the model $\mathcal{M}$, there is a tradeoff in that smaller models generally lead to narrower confidence intervals, but also may be poorer approximations to the true data generating distribution leading to lower coverage probability than desired.

Given a model $\mathcal{M}$, the method of this section requires that one specify a formula mapping the sample into a preliminary confidence interval, which we

denote by CI_1. The formula mapping the sample into a preliminary confidence interval could be, for example,

$$CI_1 := [\bar{X} - 1.96\hat{\sigma}/\sqrt{n}, \bar{X} + 1.96\hat{\sigma}/\sqrt{n}],$$

where n is the sample size and $\hat{\sigma}^2$ is the standard unbiased estimate of the variance. Under certain distributions in the model $\mathcal{M}$, this preliminary confidence interval may fail to have the desired coverage probability (e.g. at least 95% coverage). To remedy this, we will uniformly inflate the widths of these intervals by the smallest factor a such that the coverage probability is at least 95% for all distributions in the model $\mathcal{M}$. That is, we will find the smallest inflation factor $a > 0$ such that for all distributions $P \in \mathcal{M}$, the confidence interval resulting from multiplying the width of CI_1 by a has at least 95% coverage.

We illustrate the above method using an example motivated by the cluster sampling survey of mortality in Iraq of Burnham et al. (2006) that we fully discuss in Section 9. In this example, there are 47 randomly chosen clusters; we denote the number of deaths due to violence in each of these clusters by the random variables $X_1, \ldots, X_{47}$. In the example of this section, these random variables are treated as i.i.d. It is also assumed, based on prior studies, that the random counts X_i are all within the range 0 to 52. We propose using the method of this section, with statistical model $\mathcal{M}$ being the set of negative binomial distributions with parameters μ, r (representing the mean and dispersion parameters, respectively) restricted to be within the set $M \times R = [0, 52] \times [0.01, 40]$; we also truncate this distribution, in that we do not allow any values above the threshold 52, the assumed upper bound on the number of deaths per cluster. We refer to the negative binomial distribution, with values truncated at 52, as the "truncated negative binomial distribution." This set of distributions contains (approximately) Poisson distributions, but is much richer in allowing "overdispersion." (For more details, see McCullagh and Nelder (1998).)

The method of this section involves specifying a formula mapping the sample into a preliminary confidence interval CI_1, and then inflating the width of this confidence interval by the smallest multiplicative factor a that guarantees at least 95% coverage for all distributions in $\mathcal{M}$. A simple choice for the preliminary confidence interval CI_1 would be $[\bar{X} - 1.96\hat{\sigma}/\sqrt{47}, \bar{X} + 1.96\hat{\sigma}/\sqrt{47}]$, where $\hat{\sigma}^2$ is the standard unbiased estimate of the variance. However, as proved in Appendix 2, at sample size $n = 47$, no inflation factor a applied to this preliminary confidence interval is large enough to guarantee at least 95% coverage for all distributions in $\mathcal{M}$. The reason is that $\mathcal{M}$ contains distributions with non-zero mean and for which the probability that the sample consists of all

0's is greater than 0.05; in such cases, $\hat{\sigma} = 0$ and so the preliminary confidence interval CI has 0 width. Thus, no amount of inflation can ensure 95% coverage for all distributions in $\mathcal{M}$. To remedy this problem, we use a modified version of $\hat{\sigma}$ that is always greater than a fixed constant, in our preliminary confidence interval. More precisely, letting $\hat{\sigma}' := \max(\hat{\sigma}, d)$, we define our preliminary confidence interval to be

$$CI_1 := [\bar{X} - 1.96\hat{\sigma}'/\sqrt{47}, \bar{X} + 1.96\hat{\sigma}'/\sqrt{47}], \tag{13}$$

for some value $d > 0$. There is a tradeoff in choosing d, in that larger values of d lead to smaller inflation factors a, but also lead to larger values of $\hat{\sigma}'$. We discuss this tradeoff and the choice of d in Appendix 2.

For any choice of the lower bound d on $\hat{\sigma}'$, we can approximate the minimum inflation factor $a > 0$ such that for all distributions $P \in \mathcal{M}$, the confidence interval resulting from multiplying the width of CI_1 defined in (13) by a has at least 95% coverage. One way to do this is to first select a grid S of parameter choices that is as dense as possible (within the limits of available computing time required for the algorithm given in Appendix 2) in the set of parameters $M \times R$ defining the model $\mathcal{M}$. We give an example of such a set S in Appendix 2. Next, one approximates, for each choice of parameters $(\mu, r) \in S$, the minimum inflation factor $a_{\mu,r}$ for which the resulting confidence interval (for the mean) has 95% coverage when data is generated from the truncated negative binomial distribution with parameters (μ, r). This can be done by generating, say, 100,000 sets of 47 Monte Carlo draws from this truncated negative binomial distribution with parameters (μ, r), and then calculating the resulting value of the preliminary confidence interval CI_1 defined in (13) for each set; one then finds the smallest $a' > 0$ such that inflating each preliminary confidence interval by a' results in 95% of the 100,000 sets containing the true mean of the truncated negative binomial distribution with these parameters; we denote this value by $a_{\mu,r}$. We give an algorithm to approximate $a_{\mu,r}$, which is more efficient than the simple method just described, along with R code, in Appendix 2. Lastly, one sets the inflation factor a to be the maximum of $a_{\mu,r}$ over $(\mu, r) \in S$.

Using $d = 10$ in the definition of $\hat{\sigma}'$, the above procedure yields an approximate inflation factor $a = 2.39/1.96 = 1.22$. Thus, we need to inflate the width of the preliminary confidence interval (13) by 1.22, yielding the final confidence interval formula:

$$CI := [\bar{X} - 2.39\hat{\sigma}'/\sqrt{47}, \bar{X} + 2.39\hat{\sigma}'/\sqrt{47}]. \tag{14}$$

Now, given the actual data $X_1, \ldots, X_{47}$ containing the death counts in the 47 clusters of the Iraq sample survey of (Burnham et al., 2006) described in

Section 9, we calculate $\bar{X}$, $\hat{\sigma}$, and then $\hat{\sigma}' = \max(10, \hat{\sigma})$. We substitute these values into (14) to get our confidence interval for the population mean. In Section 9, we give this confidence interval and compare to confidence intervals generated by standard methods and the other methods described in this paper.

We note that simply using the parametric bootstrap with the above negative binomial model $\mathcal{M}$ will not lead to even approximately 95% coverage for all distributions in $\mathcal{M}$. This is due to the fact that $\mathcal{M}$ contains distributions with non-zero mean for which the probability that the sample consists of all 0's is greater than 0.05. When the sample consists of all 0's, the parametric bootstrap will produce a confidence interval centered at 0 with 0 width, leading to less than 95% coverage. The nonparametric bootstrap has a similar problem. The algorithm given in Appendix 2 is quite general, in that it allows any type of initial confidence interval; one could use it, for example, to compute the inflation factor required for $\mathcal{M}$ when the initial confidence interval CI_1 is based on the bootstrap (with a minimum width d as discussed above to avoid the problem of all data being 0's).

7 Comparison of Confidence Intervals Based on Different Methods

We compare the confidence intervals based on the different methods we have proposed. These include methods based on Bernstein's inequality, Bennett's inequality, and Hoeffding's inequality, as well as the method based on computing worst-case tail bounds for parametric families given in the previous section. First, we compare the assumptions that each method requires in order to guarantee it gives correct coverage, and then we compare the widths of confidence intervals from these methods. Naturally, the methods requiring stronger assumptions about the data generating distribution produce narrower confidence intervals. We also compare the assumptions and confidence interval widths for our proposed methods to those for normal-based methods.

7.1 Assumptions Required for Different Methods

We first compare the assumptions required for the methods we have proposed. Table 1 lists the assumptions required for the methods developed in this paper.

The first assumption listed in the table is that an accurate bound W is known for the maximum value. More precisely, for Bernstein's and Bennett's inequalities, a bound W must be known for the maximum absolute deviation from the mean; Hoeffding's inequality requires something slightly different:

Table 1: Assumptions Required for Various Methods of Constructing Confidence Intervals for the Population Mean. An 'X' indicates the corresponding assumption is required for that method. "Parametric" refers to the method from Section 6 that is correct when the data generating distribution belongs to a known parametric model.

	Known upper bound on max. value	Known upper bound on variance	Distribution from known model	Sample mean approx. normal
Hoeffding	X			
Bernstein	X	X		
Bennett	X	X		
Parametric			X	
Normal				X

an accurate bound W on the maximum value that could be observed; that is, $\forall i, P(|X_i| \leq W) = 1$. The second assumption in the table is that an accurate upper bound v on the variance is known. It was discussed how such bounds may be approximated in Section 4. The third assumption, required by the method from Section 6, is that the data generating distribution is within a known family of distributions. The fourth assumption is that the sample mean is approximately normally distributed.

7.2 Comparison of Widths of Confidence Intervals from Different Methods

We now compare the widths of 0.95 confidence intervals generated from the various methods discussed above. We consider the simple case of n i.i.d., mean 0, random variables $X_1, \ldots, X_n$, where it is known that $P(|X_1| \leq 1) = 1$ and $P(|X_1 - EX_1| \leq 1) = 1$. Thus, $W = 1$ is an upper bound on the maximum absolute deviation of X_1 as well as an upper bound on the maximum absolute value of X_1. Given $W = 1$, the widths of confidence intervals for the different methods we consider depend on the sample size and either the estimated standard deviation (for the method of Section 6 or normal-based methods) or a known upper bound on the standard deviation (for methods based on Bernstein's or Bennett's inequalities) of the data generating distribution. For simplicity, we assume that methods based on Bernstein's or Bennett's inequalities use the estimated standard deviation as upper bounds on the true (unknown) standard deviation. If these methods use a larger number as upper bound for the standard deviation, then their corresponding confidence intervals will have

even wider widths than depicted below.

In Figure 1 we fix the sample size at $n = 47$, in anticipation of the data analysis in Section 9 in which 47 clusters are randomly selected. We vary the estimated standard deviation s from 0 to 1, and present the widths of confidence intervals based on Hoeffding's inequality, on Bennett's inequality, on Bernstein's inequality, and on the exact parametric method from Section 6. We also include a comparison with the widths of confidence intervals based on the normal approximation. Here, we consider the normal-based interval

$$\left(\bar{X} - 1.96s/\sqrt{n}, \bar{X} + 1.96s/\sqrt{n}\right), \tag{15}$$

at sample size n. Thus, the confidence interval width that we associate with the normal-based method at sample size 47 is $2 * 1.96 * s/\sqrt{47}$.

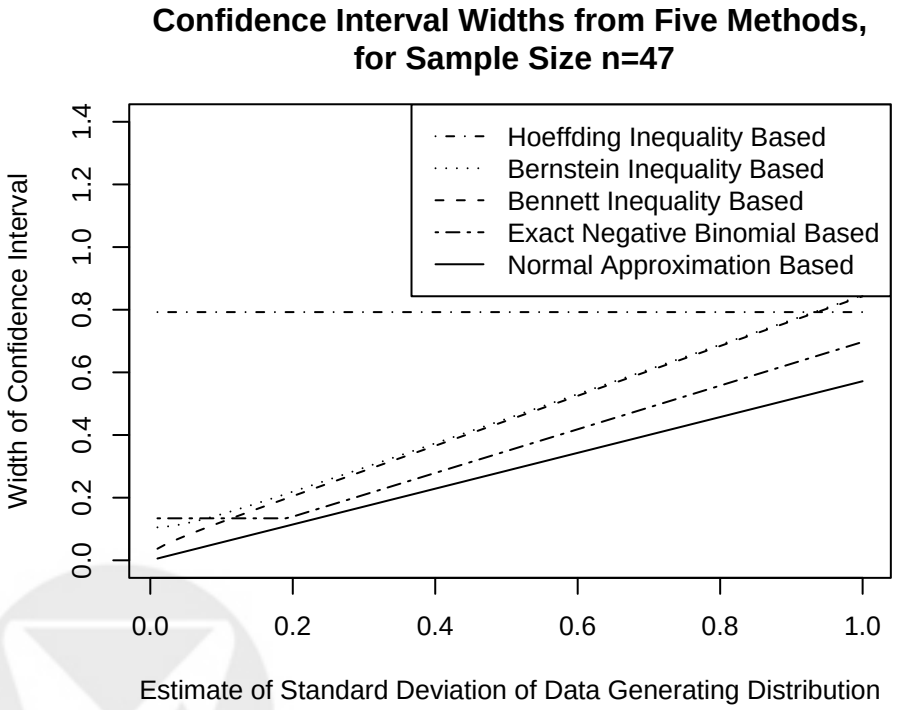

Figure 1: Confidence intervals for the mean of 47 i.i.d. random variables, all of which are known to be bounded by 1 ($-1 \leq X_i \leq 1$). The horizontal axis varies the estimated standard deviation of the data generating distribution. The width of the normal-based method is computed from (15).

First, as is expected, the larger the estimated standard deviation of the data generating distribution, the wider the resulting confidence intervals, for

all methods except that based on Hoeffding's inequality. (Hoeffding's inequality does not use any information on the standard deviation of the data generating distribution.) Confidence intervals based on Bennett's inequality and Bernstein's inequality are similar, except for very small values of the estimated standard deviation, where Bennett's does much better. Both Bernstein-based and Bennett-based methods give wider confidence intervals than the normal-based method, by a multiplicative factor that ranges from about 1.5 to 1.8 (except for estimated standard deviations < 0.2, where this factor is larger). The parametric method from Section 6 also gives wider confidence intervals than the normal-based method, but by a multiplicative factor about 1.2 (except for estimated standard deviations < 0.2, where this factor is larger); we note that our parametric method can be tuned to give tighter confidence intervals, if one is able to pre-specify (e.g. based on prior studies) a good approximation to the standard deviation–this is discussed in Appendix 2. It is important to keep in mind that under much weaker assumptions than required for normal-based methods, all the methods represented in Figure 1 except the normal-based method have coverage at least 0.95.

Figure 2 gives the same comparison as Figure 1, except sample size $n = 200$ is used instead of $n = 47$. All methods give narrower confidence intervals, by about a factor of 2. Both Bernstein-based and Bennett-based methods give wider confidence intervals than the normal-based method, by a multiplicative factor that ranges from about 1.4 to 1.6 (except for estimated standard deviations < 0.2, where this factor is larger). For the parametric method from Section 6, the ratio of confidence interval widths compared to the normal-based method is about 1.1 (except for estimated standard deviations < 0.2, where this ratio is larger).

8 Extension to General Asymptotically Linear Estimators

In this section, we give a method for constructing confidence intervals for more general parameters than the population mean. The method is based on approximating the distribution of an estimator of the parameter of interest using the influence curve for that estimator. Let θ_n be an estimator of a parameter θ_0 based on n i.i.d. observations $X_1, \ldots, X_n$, and assume that the

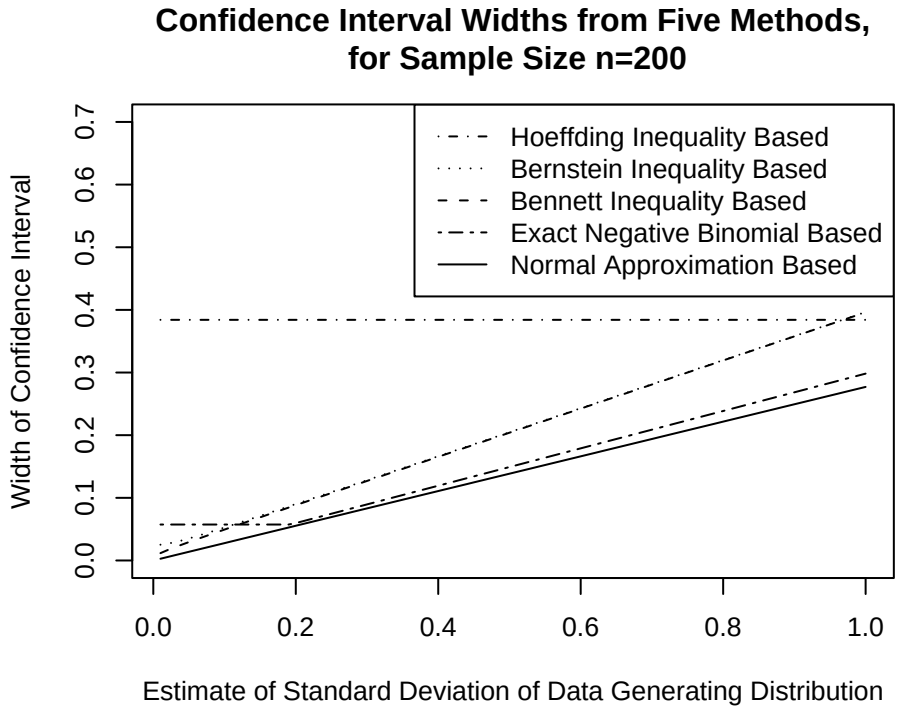

Figure 2: Confidence intervals for the mean of 200 i.i.d. random variables, all of which are known to be bounded by 1 ($-1 \leq X_i \leq 1$). The horizontal axis varies the estimated standard deviation of the data generating distribution.

estimator θ_n is asymptotically linear[3] with influence curve $IC(X_i)$:

$$\theta_n - \theta_0 = \frac{1}{n}\sum_{i=1}^{n} IC(X_i) + o_P(1/\sqrt{n}).$$

This includes a rich class of estimators, such as least squares estimators of coefficients of linear regression models, maximum likelihood estimators of coefficients of more general regression models (including logistic regression models), and standard estimators of coefficients of Cox proportional hazards models, to name a few. Standard asymptotic confidence intervals rely on this first order approximation and the central limit theorem applied to $\frac{1}{\sqrt{n}}\sum_i IC(X_i)$. We suggest that one could construct, e.g. Bernstein type confidence intervals by

[3]See van der Vaart (1998) for an overview of asymptotically linear estimators.

using the approximation

$$P\left(|\theta_n - \theta_0| > q/\sqrt{n}\right) \approx P\left(|\frac{1}{n}\sum_{i=1}^{n} IC(X_i)| > q/\sqrt{n}\right),$$

and applying Bernstein's inequality to the latter empirical mean of i.i.d. random variables $IC(X_i)$. This results in the same confidence interval as in Theorem 1 presented in Section 4, but with the role of $(X_i - \mu)$ replaced by $IC(X_i)$. Our methods based on the other tail bound inequalities from Section 5, and our parametric-based methods from Section 6 can be applied similarly.

The above method allows for constructing conservative confidence intervals for parameters that are more challenging, in terms of statistical inference, than the sample mean. One example is causal effect parameters such as those described in (Bembom et al., 2008), for which data relevant to the question of interest is sparse. As discussed in (Bembom et al., 2008), such parameters may have influence curves that are skewed and so $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} IC(X_i)$ will not be approximately normal. In this case, using the methods of this paper can give more conservative confidence intervals than one would obtain by using e.g. normal-based methods.

9 Data Analysis Example: Estimating Iraq Mortality

We illustrate the above methods by applying them to the data used in "Mortality after the 2003 invasion of Iraq: A cross-sectional cluster sample survey", by Burnham et al. (2006). We first summarize the methods and results of Burnham et al. (2006). We then present Bernstein/Bennett-based confidence intervals, and then confidence intervals based on the parametric method from Section 6.

9.1 Summary of Design and Results from Burnham et al. (2006)

We briefly summarize the survey design and the main result for deaths due to violence given in (Burnham et al., 2006). The survey design involved dividing up Iraq in 18 Governorates. 50 clusters were allocated systematically among the Governorates in proportion to population size. The clusters allocated to each area were distributed within constituent administrative units, proportional to population size. In each administrative unit that was allocated a

cluster, approximately 40 neighboring households were randomly selected (see (Burnham et al., 2006) for the procedure used for this selection). These households were questioned regarding deaths of household members during the period being studied. Numbers of deaths and cause of death were recorded. Two clusters were not visited due to miscommunication and one was not visited due to insecurity in that area; we focus on the 47 remaining clusters. Denote the observed counts of violent deaths in each cluster by $x_1, \ldots, x_{47}$. The sample mean $\bar{x} = 6.43$ and the sample variance $\hat{\sigma}^2 = 69.16$. These data were mapped into an estimate of the mean number of deaths due to violence in all of Iraq in the time period under consideration. This was done by calculating mortality rates using log-linear regression models, and then scaling up using an estimate for the total population of Iraq. The final estimate was 601,027 deaths due to violence between March 2003 and July 2006. The 0.95 confidence interval of Burnham et al. for this point estimate was

$$\text{Burnham et al. (2006) Confidence Interval: } (426,369 - 793,663).$$

It was calculated using robust variance estimators for log-linear regression models, and ended up being similar to confidence intervals based on the nonparametric bootstrap (Burnham et al., 2006).

9.2 Confidence Intervals for Number of Violent Deaths Based on Bernstein's Inequality and Bennett's Inequality

Below, we calculate confidence intervals for the number of deaths due to violence using this same set of 47 counts as above, but applying the Bernstein-based and Bennett-based methods described in this paper. Our motivation is that since we are confronted with a small sample size for a potentially highly skewed probability distribution, standard model-based or bootstrap-based methods may not lead to accurate inference.

The following Bernstein-based 0.95 confidence interval heavily relies on the assumption that the sample represents 47 independently (but not necessarily identically) distributed geographical clusters of households. Furthermore, within every Governorate sampled, each household must have an equal chance of being selected in any cluster allocated to that Governorate. We emphasize that our method cannot correct any bias in the sampling procedure; minimizing potential bias is a major challenge when doing cluster surveys in conflict areas such as Iraq. For example, sampling bias could result from choosing clusters located close to main streets (called "main street bias"), from accurate

estimates of population density in different areas being unavailable in conflict settings, from the sequence of households selected within a cluster being influenced by necessary choices of the survey team on the ground (which can unintentionally lead to selection bias), from non-response bias, and from cluster size being potentially correlated with the level of violence in an area. Since the purpose of this section is to demonstrate the construction of Bernstein-based confidence intervals in a concrete survey sampling example, we do not address the issue of sampling bias. For further discussion of issues related to the survey of Burnham et al. (2006) that are not considered here, see e.g. (von Schreeb et al., 2007; Hicks, 2007; Burnham et al., 2007; Brownstein and Brownstein, 2008; Alkhuzai et al., 2008).

As in Section 2.2, let $X_1, \ldots, X_{47}$ denote the random variables representing the number of deaths due to violence in each of the 47 randomly chosen clusters. We are concerned with using this data set to obtain an estimate of $\mu = \sum_{k=1}^{K} w(k)\mu(k)$, where for each Governorate (sub-area) k, $w(k)$ is the proportion of the total population of Iraq in that Governorate, and $\mu(k)$ is the mean number of deaths due to violence in a randomly chosen cluster in that Governorate. An estimate and a corresponding confidence interval for μ can be scaled up to an estimate and confidence interval for the number D of violent deaths in all of Iraq during the studied post-invasion period, as described in Section 2.2. The scaling factor, as derived in Section 2.2, is the (estimated) total population of Iraq divided by the estimated mean number of individuals per cluster, which equals $(27139584)/(12801/47) = 99645.38$.

Our estimator for μ is $\sum_{k=1}^{K} w(k)\bar{X}_k$, where $\bar{X}_k$ is the sample mean within the kth Governorate. As shown in Section 2.2, when clusters are allocated proportional to population size in each sub-area, which was the case in this study design, then $\sum_{k=1}^{K} w(k)\bar{X}_k = \frac{1}{n}\sum_{i=1}^{n} X_i$, where $n = 47$ is the total number of clusters. In order to construct Bernstein-based confidence intervals as described in Section 4, we need to specify an upper bound W on the maximum possible value of $|X_i - EX_i|$ and an upper bound σ^{*2} on the variance of the sum $\sum_{i=1}^{47} X_i$ used in our estimator for μ. W should, at the very least, be chosen larger than the maximum difference between all observed counts in this data set and μ, and the maximum difference of all observed counts in a previous study in Iraq (Roberts et al., 2004) and μ. The maximum count in this data set is 35 and the maximum in the previous study is 52. Therefore, $W = 50$ seems a reasonable choice. We next consider how to select an appropriate bound σ^{*2} on the variance of $\sum_{i=1}^{47} X_i$.

We base our choice of σ^{*2} on $n\hat{\sigma}^2$, for $\hat{\sigma}^2$ the standard unbiased estimate of the variance treating $\{X_i\}$ as if they were i.i.d. As discussed in Section 2.2, $n\hat{\sigma}^2$ has mean that is greater than or equal to the actual variance of $\sum_{i=1}^{47} X_i$,

with equality when X_i all have the same mean. We therefore use $n\hat{\sigma}^2$ as our approximate upper bound σ^{*2}. The Bernstein-based confidence interval, based on using these choices of W and σ^{*2} in Theorem 2, is

$$\text{Bernstein-based Confidence Interval: } (157,123 - 1,124,316).$$

The R-code for this analysis is provided in Appendix 3. Using the same analysis as above, but instead relying on Bennett's inequality instead of Bernstein's inequality gives similar confidence intervals, but about 9% smaller.

As a sensitivity analysis, we look at the effect of using larger or smaller values for the upper bound σ^{*2} used in Bernstein's inequality. One choice for how much to vary σ^{*2} can be based on a rough estimate of the variability of the estimator we used for it. If we treat the sample as if it were i.i.d. copies of a random variable X, then the standard error of $\hat{\sigma}^2$ could be estimated using the fact that $\hat{\sigma}^2$ would be an asymptotically linear estimator of the variance of X, and would have influence curve $IC(X) = X^2 - EX^2 - 2EX(X - EX)$; so, the variance of $\hat{\sigma}^2$ could be approximated to first order with $\text{VAR}(IC(X))/47$. The standard error in $\hat{\sigma}^2$, estimated in this manner using moments estimated from the data, is 20.5. The resulting confidence interval, based on adding one estimated standard error to $\hat{\sigma}^2$, is then:

$$(115,458 - 1,165,981).$$

Similarly, to be even more conservative, we can add two estimated standard errors to $\hat{\sigma}^2$, resulting in the confidence interval:

$$(76,277 - 1,205,163).$$

We provide R-code in Appendix 3 for the above calculations.

9.3 Confidence Interval for Number of Violent Deaths Based on Method Using Parametric Models from Section 6

We apply the method described in Section 6 to the data in (Burnham et al., 2006). This method involves first specifying a statistical model $\mathcal{M}$ for the (unknown) distribution of the random variable X, which represents the number of deaths due to violence in a randomly chosen cluster according to the sampling methodology used in (Burnham et al., 2006). For simplicity and ease of demonstration, we assume the cluster counts are i.i.d., even though the actual sampling mechanism (described above) involved partitioning the

total area into subareas and sampling proportional to population within each subarea.

We define the model $\mathcal{M}$ to be the set of negative binomial distributions with parameters μ, r (representing the mean and dispersion parameters, respectively) restricted to be within the set $M \times R = [0, 52] \times [0.01, 40]$, and truncated to have values at most 52 (the assumed maximum possible count in any cluster). The method of Section 6 outputs a confidence interval for the mean that has correct coverage probability whenever the data is generated according to a distribution in the model $\mathcal{M}$. We decided to let the model $\mathcal{M}$ be a set of negative binomial distributions, since the more restrictive Poisson model did not fit the data (the mean of the data was 6.43 and the sample variance was $\hat{\sigma}^2 = 69.16$, while in a Poisson model the mean and variance are equal), and since the negative binomial family includes highly skewed distributions as well as standard distributions such as the Poisson.

The method in Section 6 requires specifying a formula for a preliminary confidence interval, which we define to be

$$CI_1 := [\bar{X} - 1.96\hat{\sigma}'/\sqrt{47}, \bar{X} + 1.96\hat{\sigma}'/\sqrt{47}], \tag{16}$$

where $\hat{\sigma}' := \max(\hat{\sigma}, 10)$. (See Section 6 for why we need to use this modified version of $\hat{\sigma}$.)

We then approximate the minimum inflation factor $a > 0$ such that for all distributions $P \in \mathcal{M}$, the confidence interval resulting from multiplying the width of CI_1 defined in (16) by a has at least 95% coverage. The algorithm and R code for this computation are given in Appendix 2. The result is an approximate inflation factor of $a = 2.39/1.96 = 1.22$, which leads to the final confidence interval formula:

$$CI := [\bar{X} - 2.39\hat{\sigma}'/\sqrt{47}, \bar{X} + 2.39\hat{\sigma}'/\sqrt{47}]. \tag{17}$$

Substituting the observed values from the Iraq data ($\bar{X} = 6.43$ and $\hat{\sigma}' = \max(\sqrt{69.16}, 10) = 10$) into the above display, and scaling up by the factor 99645.38 which maps the mean count to the total number of deaths due to violence in Iraq during the study period, yields the following confidence interval:

Parametric Method-Based Confidence Interval $(293,339 - 988,101)$.

This confidence interval has much smaller width than those generated by the Bernstein-based method, but has larger width than the confidence interval $(426,369 - 793,663)$ of Burnham et al. (2006) that was based on log-linear models.

10 Discussion

In this paper, we describe a framework for construction of more robust confidence intervals for the mean of a distribution, tailored to small sample sizes. The first set of methods in our framework relies on a tail bound, such as Bernstein's inequality, Bennett's inequality, or Hoeffding's inequality. As described in Section 5, improvements on any of these inequalities can be immediately incorporated into our framework, resulting in improved confidence intervals for small sample sizes. Because the widths of confidence intervals produced by this first set of methods can be much larger than those of normal-based methods, we presented a second set of methods. This second set of methods produces confidence intervals with shorter widths than the first set of methods, and has correct coverage[4] for all sample sizes if the data generating distribution is in a given parametric model.

We now lay out a template for future research aimed at constructing confidence intervals tailored to small sample sizes. We think the following research questions are of primary interest:

1. **Sharper Tail Bounds:** An important open question is whether the gap between normal-based and Bennett-based confidence interval methods as discussed in Section 7 is due to the Bennett bounds being loose or not. It as an open problem to resolve this, either by sharpening the Bennett bounds, or by exhibiting a class of counterexamples showing that Bennett-type bounds are close to sharp. A similar open question applies to the Hoeffding bounds.

   One approach to this problem is to select a parametric model and a fixed sample size, and then compute approximate tail bounds on the sample mean using statistical software. Such tail bounds can be found by approximating

   $$\sup_{\theta \in \Theta'} \left(\inf\{q \geq 0 : P_\theta \left(|\bar{X} - E_\theta X| \leq q \right) \geq 0.95\} \right),$$

   where Θ' is a subset of the parameter space meeting a specified set of restrictions (such as variance at most σ^2), P_θ is the probability distribution at θ, E_θ is the mean with respect to θ, and $\bar{X}$ is the sample mean (at a fixed sample size). This can be approximated using a method similar to that described in Section 6 and in Appendix 2. If the bounds of

[4] The precise coverage probability will depend on the coarseness of the grid used in the algorithm given in Appendix 2.

an inequality are closely matched by the actual bounds of a parametric family, this would show the inequality is sharp.

2. **Applying the Method of Section 6 to Other Parametric Families and Starting with Other Initial Confidence Intervals:** The parametric-based method of Section 6 takes as input a given parametric model and an initial confidence interval method (such as a standard normal-based method). It then computes the smallest inflation factor by which the initial confidence interval would have to be expanded in order to guarantee 0.95 coverage for every distribution in the model. This was done for the negative binomial model with parameters restricted as described in Section 6, and further explored in Appendix 2. This general method could be applied using other types of initial confidence intervals, for example, using bootstrap-based initial confidence intervals, and determining how much these would have to be inflated to guarantee 0.95 coverage for every distribution in a given model. Also, other models than the negative binomial could be considered. One goal would be, for a given parametric model, to find the type of initial confidence interval that leads to final (inflated) confidence intervals having the smallest widths.

3. **Realistic, Stronger Assumptions:** If one could specify assumptions that are often true in practice under which one can prove substantially stronger tail bounds, one could immediately obtain narrower confidence intervals with guaranteed probability of coverage under those assumptions. These assumptions would presumably be based on prior studies or domain-specific knowledge of the data generating distribution.

4. **Robust Parameters:** It would be useful to obtain methods for constructing exact confidence intervals for more robust parameters than the mean, such as the trimmed mean. We note that one can construct exact confidence intervals for the median, using the order statistics $X_{(1)}, \ldots, X_{(n)}$ from an i.i.d. sample $X_1, \ldots, X_n$, based on the binomial distribution, as follows:
Let k be the largest integer such that $\sum_{i=0}^{k-1} \binom{n}{i}/2^n \leq 0.025$; then, for any data generating distribution, $(X_{(k)}, X_{(n-k+1)})$ is a 95% confidence interval for the median. This follows since the interval $(X_{(k)}, X_{(n-k+1)})$ fails to contain the median if and only if either (a) fewer than k of $X_1, \ldots, X_n$ are less than or equal to the median or (b) fewer than k of $X_1, \ldots, X_n$ are greater than or equal to the median. By the definition of median, the probability that X_1 is less than or equal to the median is at least 1/2;

similarly, the probability that X_1 is greater than or equal to the median is at least $1/2$. Thus, using the fact that the X_i are i.i.d., we have the probability that (a) or (b) occurs is at most $2\sum_{i=0}^{k-1}\binom{n}{i}/2^n$, which is less than or equal to 0.05 by construction. Thus, $(X_{(k)}, X_{(n-k+1)})$ has 95% coverage for the median.

Appendix 1: Proofs of Claims from Section 3

We prove the claims in the three examples from Section 3.

Example 1

First, consider the construction in Example 1 from Section 3, which we repeat here: Let δ be a number in $(0,1)$. Let A take value 0 with probability $1-\delta$ and take value 1 with probability δ. Let Y be a zero-mean, normal random variable with variance τ^2. Assume A and Y are independent. Let $X = A+Y$. Assume we observe n i.i.d. copies of X, denoted by $X_1, \ldots, X_n$. Denote the corresponding values of A and Y for each such copy by $(A_1, Y_1), \ldots, (A_n, Y_n)$.

Theorem 3 *For any $\epsilon > 0$, for any sample size n, there is a $\delta > 0$ and $\tau^2 > 0$ such that the 95% confidence intervals based on the normal distribution, Student's t-distribution, or nonparametric bootstrap will contain the mean of X with probability at most ϵ.*

Proof: The probability that for all i, $A_i = 0$, is $(1-\delta)^n$. We can choose $\delta > 0$ small enough so that this probability is at least $1 - \epsilon/2$. Consider the normal-based confidence interval with right hand endpoint $\bar{X} + 1.96\hat{\sigma}/\sqrt{n}$. We can choose τ^2 small enough that $P(\bar{X} + 1.96\hat{\sigma}/\sqrt{n} \geq \delta | \forall i, A_i = 0) < \epsilon/2$, where $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ is an unbiased variance estimator. This follows since, letting $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$, we have

$$
\begin{aligned}
& P\left(\bar{X} + 1.96\hat{\sigma}/\sqrt{n} \geq \delta | \forall i, A_i = 0\right) \\
&= P\left(\bar{Y} + 1.96\sqrt{\frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2}/\sqrt{n} \geq \delta | \forall i, A_i = 0\right) \\
&= P\left(\bar{Y} + 1.96\sqrt{\frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2}/\sqrt{n} \geq \delta\right) \\
&\leq E\left(\bar{Y} + 1.96\sqrt{\frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2}/\sqrt{n}\right)^2 /\delta^2
\end{aligned}
$$

$$= (\tau^2/n + 1.96^2\tau^2/n)/\delta^2 \tag{18}$$

where the third line follows since $\{Y_i\}$ and $\{A_i\}$ are independent, the fourth line follows from Chebychev's inequality, and the last line follows from the fact that $\bar{Y}$ and $\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$ are independent due to Y_i being normally distributed.

Therefore, we can choose τ^2 sufficiently smaller than δ such that the right hand side of the last equation is less than $\epsilon/2$. Thus, the probability that the normal-based confidence interval contains the true mean (δ) is at most

$$P(\bar{X} + 1.96\hat{\sigma}/\sqrt{n} \geq \delta | \forall i, A_i = 0)P(\forall i, A_i = 0) + 1 - P(\forall i, A_i = 0) < \epsilon.$$

We now focus on sample size $n = 50$. Then for $\delta = 0.01$ and $v = 0.032$, we have, based on a simulation in R, that the coverage probability of the normal-based 0.95 confidence interval is ≈ 0.64. (The simulation involved generating 50 copies of X and calculating the corresponding normal-based confidence interval. This was repeated $100,000$ times, and the fraction of iterations for which the interval contained the true mean δ was recorded.)

A similar argument as above, except using the 0.975 quantile of the t-distribution with $n-1$ degrees of freedom in place of 1.96, gives that for any $\epsilon > 0$, we can choose δ and τ such that the coverage probability using Student's t-distribution is at most ϵ. For $n = 50$, $\delta = 0.01$, and $\tau = 0.032$, we have this coverage probability is ≈ 0.65, based on a simulation in R.

Lastly, we show δ and τ can be chosen so that confidence intervals based on the nonparametric bootstrap percentile method will contain the true mean with probability at most ϵ. First, choose $\delta > 0$ small enough such that the following event has probability at least $1 - \epsilon/4$:
Event A: $\forall i, A_i = 0$.
Next, choose $\tau > 0$ such that the following three events each have probability at least $1 - \epsilon/4$:
Event B: $\bar{Y} < \delta/2$.
Event C: $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 < \delta^2 n/(-16 \ln 0.025)$
Event D: $\max_i(Y_i - \bar{Y}) < 3\delta n/(-8 \ln 0.025)$

Now, we show that conditioning on the intersection of the above events, the nonparametric bootstrap distribution has 0.975 quantile less than δ; this implies that on the intersection of these events, the nonparametric bootstrap percentile method confidence interval will have right-limit less than δ, so will fail to contain the true mean (δ). Denote the observed sample variance $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ by v. Denote the observed maximum absolute deviation from

the mean, $\max_i(Y_i - \bar{Y})$, by W. Here v and W are random, depending on $Y_1, \ldots, Y_n$.

Let $Y_1^*, \ldots, Y_n^*$ be a set of bootstrap replicates of $Y_1, \ldots, Y_n$; that is, let $Y_1^*, \ldots, Y_n^*$ be i.i.d. draws (with replacement) from the empirical distribution of the observations $Y_1, \ldots, Y_n$. Then it follows from Bernstein's inequality that

$$\begin{aligned} & P\left(\frac{1}{n}\sum Y_i^* \geq \delta | A \cap B \cap C \cap D\right) \\ \leq \; & P\left(\frac{1}{n}\sum Y_i^* - \bar{Y} > \delta/2 | A \cap B \cap C \cap D\right) \\ \leq \; & E\left(\exp[-n^2\delta^2/(8nv + 4Wn\delta/3)] | A \cap B \cap C \cap D\right) \\ < \; & 0.025 \end{aligned}$$

where the first inequality follows by the definition of Event B; the second inequality follows from Bernstein's inequality since the bootstrap replicates are i.i.d. conditioned on $A_1, \ldots, A_n, Y_1, \ldots, Y_n$; and the third inequality follows from the definition of Events C and D. Thus, on $A \cap B \cap C \cap D$, the percentile bootstrap 0.95 confidence interval will have right endpoint less than δ, thus failing to contain the true mean. Since δ and τ were chosen so that $P(A \cap B \cap C \cap D) > 1 - \epsilon$, the percentile bootstrap 0.95 confidence interval contains the true mean with probability at most ϵ.

We now focus on sample size $n = 50$. Then for $\delta = 0.01$ and $v = 0.032$, we have, based on a simulation in R, that the coverage probability of the nonparametric bootstrap percentile method 0.95 confidence interval is ≈ 0.63.

Example 2

Now consider Example 2 from Section 3. The setup was the same as in Example 1, except that here we assume the variance of X, denoted by σ^2, is known.

We show the following, based on numerical calculations using the statistical software R:
For any sample size $n : 1 < n < 1000000$, there is a distribution for which the normal-based confidence interval (4), in which the true standard deviation is used, contains the true mean with probability at most 0.84. The distribution is constructed exactly as in Example 1, and then choosing δ and τ as we describe below. In particular, for $n = 50$, if we choose $\delta = 0.0035$ and $\tau = 0.0001$, then the probability that (4) contains the true mean is approximately 0.839. Furthermore, for $n = 50$, the normal-based 0.95 confidence interval would have to be increased by a multiplicative factor of at least 2.13 in order for it to have correct coverage for the set of distributions we construct.

To show this, we first show for any $n : 1 < n < 1000000$, how to choose a corresponding δ small enough such that the normal-based confidence interval will fail to contain the true mean δ whenever at least one of the $A_i = 1$. Initially, let $\tau = 0$, so that X takes values in $\{0, 1\}$, with probability δ of being 1; later we will increase τ. For $n > 0$, define δ' to be the solution of

$$\delta = \frac{1}{n} - 1.96\sqrt{\delta(1-\delta)/n}.$$

That is,

$$\delta' = \left(1.96^2 + 2 - \sqrt{(1.96^2+2)^2 - 4 - 4 * 1.96^2/n}\right)/(2(n + 1.96^2)). \quad (19)$$

Choose δ to be any positive number less than δ'. Then the probability that the normal-based confidence interval contains the true mean δ is at most

$$P(\bar{X} - 1.96\sigma/\sqrt{n} \leq \delta) = P(\forall i, A_i = 0) = (1-\delta)^n. \quad (20)$$

It is straightforward to compute the right hand side of the previous display, when substituting δ' as defined above for δ, for n taking integer values between 1 and 1000000. This was done in R and it is always less than 0.839 for such n.

Thus, for all $n : 1 < n < 1000000$, δ and sufficiently small τ can be chosen such that the coverage probability for the normal-based method at nominal level 0.95 is at most 0.84.

For $n = 50$, we have $\delta' = 0.00354$; then choosing $\delta = 0.0035$, the right hand side of (20) is 0.839. Choosing $\tau = 0.0001$, we have that the coverage probability of the normal-based method using the correct value of the standard deviation σ is approximately 0.839, based on 1000000 Monte Carlo simulations of 50 draws from the distribution of this example.

Lastly, we explore by what factor the width of the normal-based confidence interval would have to be increased in order to provide 0.95 coverage for the types of distributions described in this example. That is, we consider replacing the 1.96 in (4) by a larger number so that it will have 0.95 coverage for distributions defined by $n = 50$, and various choices of δ and τ. The 1.96 would have to be replaced by at least 4.18, as we argue next.

For sample size $n = 50$, consider replacing 1.96 by a larger number K (to be determined shortly) in the definition of δ' above (19). Then substituting this δ' for δ in the right hand side of (20), we can compute the probability of coverage for the interval $(\bar{X} - K\sigma/\sqrt{n}, \bar{X} + K\sigma/\sqrt{n})$. This can be done numerically, and for $K \leq 4.18$, we have that the coverage probability is smaller than 0.95. This means that the width of the normal-based confidence interval

(4) would have to be increased by at least a factor of $4.18/1.96 = 2.13$ in order to have correct coverage probability for this family of distributions.

Example 3

Consider Example 3 from Section 3. We had constructed a random variable as follows: Let Y be a Poisson random variable with mean 1. Let A be independent of Y and take value 0 with probability $1 - \delta$ and take value v with probability δ. Let $X = A + Y$. Assume we observe n i.i.d. copies of X. We have the following theorem:

Theorem 4 *For any $\epsilon > 0$, for any sample size n, if δ is sufficiently small and v is sufficiently large, the 0.95 confidence intervals based on the Poisson distribution using the parametric bootstrap will contain the mean of X with probability at most ϵ.*

Proof: The mean of X is $1 + \delta v$. The probability that for all i, $A_i = 0$ is $(1-\delta)^n$. Choosing δ small enough so that this probability is at least $1-\epsilon/2$, and then choosing v large enough so that the probability the parametric bootstrap-based confidence interval when based on $Y_1, \ldots, Y_n$ contains $1 + \delta v$ is at most $\epsilon/2$, the theorem follows. This probability can be simulated to show that for $n = 50$, $\delta = 0.01$, and $v = 50$, we have that the probability that the parametric bootstrap-based confidence interval contains the true mean is less than 0.50.

Appendix 2: Details of Method Based on Parametric Models from Section 6, including R-code

We give the details of the method for constructing confidence intervals presented in Section 6 and used in the data example in Section 9. We first give the algorithm and then describe the resulting inflation factors and confidence interval widths. We then discuss problems with using the parametric bootstrap to construct confidence intervals for the negative binomial model we use. Lastly, we relate the method of Section 6 to a hypothesis test inversion procedure.

The negative binomial distribution parametrized by (p, r) assigns probability mass

$$\frac{\Gamma(r+k)}{\Gamma(r)k!}p^r(1-p)^k,$$

to k for each nonnegative integer k. We use the alternative parametrization (μ, r), in which $p = r/(r + \mu)$. We let our model $\mathcal{M}$ be the set of negative binomial distributions corresponding to the set of parameters $M \times R = [0, 52] \times$

$[0.01, 40]$. Recall that we let $\hat{\sigma}' := \max(\hat{\sigma}, d)$, where $\hat{\sigma}^2$ is the standard unbiased estimate for the variance and d is a pre-specified constant. (As described in Section 6, we require the lower bound d to deal with case in which the entire sample consists of 0's, which for sample size $n = 47$ and some distributions in $\mathcal{M}$ (such as for $(\mu, r) = (1, 0.01)$) occurs with probability greater than 0.05; in such cases $\hat{\sigma} = 0$, and so no inflation factor for a normal-based confidence interval would provide 0.95 coverage for such distributions.)

We now consider a grid of parameter choices in the set of parameters $M \times R$ defining the model $\mathcal{M}$. We define
$S := \{1, 2, \ldots, 52\} \times (\{0.01, 0.02, ..., 0.99, 1\} \cup \{1.5, 1.9, 2.3, \ldots, 39.6, 40\})$. This set was chosen to be as large as possible while allowing the computation described below to finish in three days of computing time; we used closer spacing for values of the dispersion parameter r less than 1, since these correspond to high skewness and the largest inflation factors were found to occur at such values of r. For each pair of parameters $(\mu, r) \in S$, we approximate the minimum inflation factor $a_{\mu,r}$ for which the resulting confidence interval (for the mean) has 95% coverage for the negative binomial distribution with parameters (μ, r) and truncated at 52, using the algorithm described next.

We now give the algorithm for computing, given any pair (μ, r), the minimum inflation factor $a_{\mu,r}$ required so that the confidence interval

$$CI := [\bar{X} - 1.96 a_{\mu,r} \hat{\sigma}'/\sqrt{47}, \bar{X} + 1.96 a_{\mu,r} \hat{\sigma}'/\sqrt{47}], \tag{21}$$

has 95% probability of containing the true mean under the negative binomial distribution with these parameters and truncated at 52. The algorithm involves first approximating the true mean $\bar{\mu}$, using 1000000 Monte Carle draws from this distribution. Next, we leverage the fact that the true mean $\bar{\mu}$ is contained in CI if and only if $\sqrt{47}|\bar{X} - \bar{\mu}|/(1.96\hat{\sigma}') \leq a_{\mu,r}$. Thus, the minimum value of $a_{\mu,r}$ for which CI contains the true mean with probability at least 95% is the minimum value of $a_{\mu,r}$ such that

$$P_{\mu,r}(\sqrt{47}|\bar{X} - \bar{\mu}|/(1.96\hat{\sigma}') \leq a_{\mu,r}) = 0.95.$$

This latter value can be approximated by randomly generating 100000 sets of 47 i.i.d. copies of a random variable with truncated negative binomial distirbution with parameters μ, r and truncated at 52, computing
$\sqrt{47}|\bar{X} - \bar{\mu}|/(1.96\hat{\sigma}')$ for each set, and then finding the 0.95 quantile of the resulting 100000 values. This is what is done in the R code at the end of this appendix.

We now give the resulting inflation factors computed by the above algorithm, for various choices of d, using the set of parameters in the grid $S :=$

$\{1, 2, \ldots, 52\} \times (\{0.01, 0.02, ..., 0.99, 1\} \cup \{1.5, 1.9, 2.3, \ldots, 39.6, 40\})$. Table 2 gives the inflation factors, for sample sizes 47 and 200, and $d \in \{1, 5, 10, 20, 25\}$. Also listed are the parameters $(\mu, r) \in S$ corresponding to the distribution that required the largest inflation factor among all distributions corresponding to parameters in S.

Table 2: For sample size $n = 47$, for various lower bounds d in the definition of $\hat{\sigma}' = \max(\hat{\sigma}, d)$, we give the inflation factor a computed by the above algorithm, as well as the "worst-case" parameters μ, r corresponding to this inflation factor.

At Sample Size 47:

d	a	μ	r
1	8.43	50	0.01
5	1.81	21	0.02
10	1.21	41	0.04
20	1.03	52	0.44
25	0.87	52	0.26

At Sample Size 200:

d	a	μ	r
1	2.63	1	0.01
5	1.23	10	0.01
10	1.07	39	0.02
20	1.02	52	0.29
25	0.88	52	0.29

We next give graphs comparing the widths of confidence intervals corresponding to different choices of d, analogous to the graphs given in Section 7. Recall that confidence interval widths, for the confidence intervals based on the method from Section 6, equal $2 * 1.96 * a \max(\hat{\sigma}, d)/\sqrt{47}$, where a is the inflation factor computed from the above algorithm. Thus, there is a tradeoff in that larger values of d increase the term $\max(\hat{\sigma}, d)$ but also lead to decreased inflation factors a (as shown in Table 2). Figures 3 and 4 correspond to sample sizes 47 and 200, respectively, and include the widths for the standard normal-based method, for comparison. In practice, d should be chosen prior to looking at the current data, and could be based on estimates of σ using

past studies. In data analysis example from Section 9, we chose $d = 10$ for our analysis somewhat arbitrarily; we note that the resulting confidence intervals for $d = 5$ would have been 24% wider.

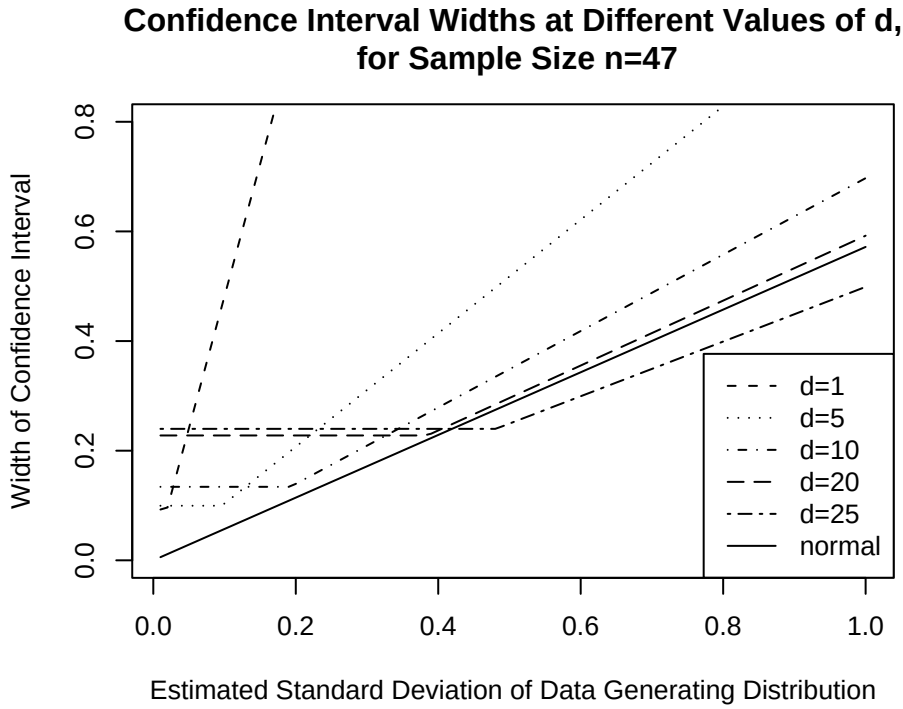

Figure 3: Confidence interval widths for the mean of 47 i.i.d. random variables, all of which are known to be bounded by 1 ($-1 \leq X_i \leq 1$), using the method from Section 6. Widths corresponding to various choices of d, the prespecified lower bound on σ', are shown. We also include, for comparison, the widths corresponding to the normal-based method, calculated from (15).

We now consider problems with using the parametric bootstrap to construct confidence intervals for our parametric model defined above. One problem is that, at sample size $n = 47$, for some distributions in our model with non-zero mean, the sample consists of all 0's with high probability. For example, the negative binomial distribution with parameters $(\mu, r) = (1, 0.01)$ assigns probability 0.954 to 0, and so the probability of all 47 data points being 0 is 0.11. When the sample consists of all 0's, the maximum likelihood estimate for (μ, r) sets $\mu = 0$, so that the parametric bootstrap will return a confidence interval centered at 0 with 0 width, thereby excluding the

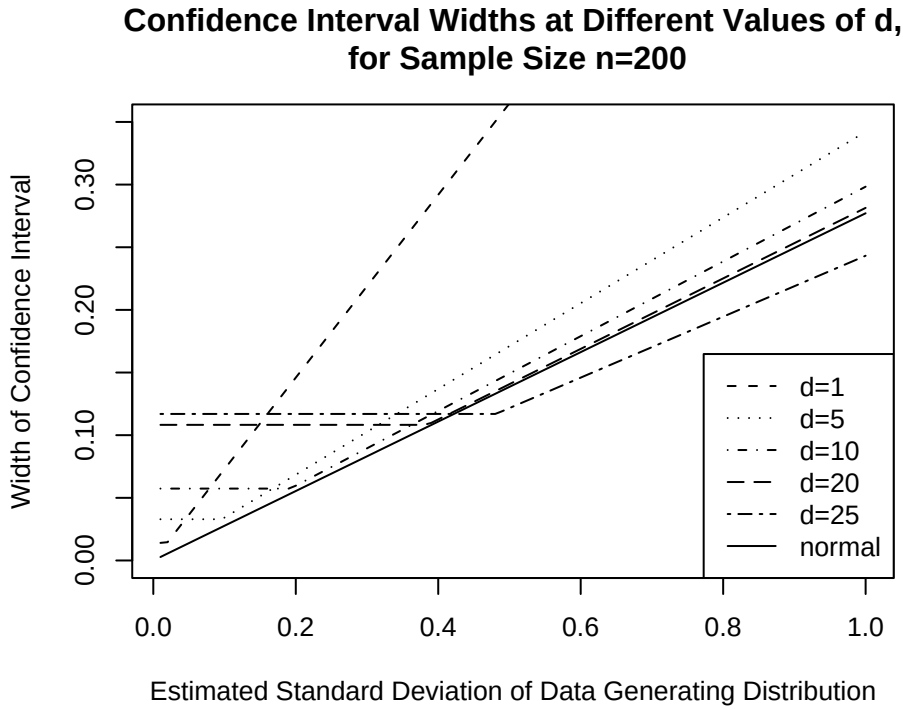

Figure 4: Confidence interval widths for the mean of 200 i.i.d. random variables, all of which are known to be bounded by 1 ($-1 \leq X_i \leq 1$), using the method from Section 6. Widths corresponding to various choices of d, the prespecified lower bound on σ', are shown. We also include, for comparison, the widths corresponding to the normal-based method, calculated from (15).

actual mean. Thus, the parametric bootstrap, without some kind of modification (such as enforcing a minimum width to all confidence intervals, as we did above for normal-based intervals) cannot be used to construct confidence intervals that have 0.95 coverage for all distributions in our model $\mathcal{M}$.

The method of Section 6 is related to a method based on inverting hypothesis tests. (See (Casella and Berger, 1990) for an overview of exact tests based on hypothesis test inversion.) This hypothesis test inversion method consists of the following steps: Based on the observed data, for each distribution $P \in \mathcal{M}$, test the null hypothesis that the mean of the data generating distribution equals the mean under P (denoted by μ_P); this can be done by checking if $|\bar{X} - \mu_P| < 1.96 a_P \hat{\sigma}'/\sqrt{n}$, where a_P is the inflation factor calculated for the distribution P as given in the algorithm above. The corresponding

confidence set, based on inverting this hypothesis test, includes all values of μ for which there exists a distribution $P \in \mathcal{M}$ for which $\mu = \mu_P$ and the corresponding hypothesis test failed to reject the null. We note that this hypothesis test inversion could be carried out using the above algorithm for our method from Section 6, but additionally storing, for each distribution (parametrized by (μ, r)), the corresponding mean and inflation factor $a_{\mu,r}$.

We now give R-code for the algorithm described in the beginning of this Appendix.

```
# Set sample size
SampleSize <- 47
# Set number of iterations in simulations
iter <- 1000
#Set value of d
d <- 5
# Set variable to hold largest value of a found so far
max_a_val <- 0
# Set truncation level (52 is assumed maximum count
# possible for Iraq study)
truncation_level <- 52
for(nb_mu in seq(1,52,length=20))
{
for(nb_disp in c(seq(0.01,1,length=100),seq(1.5,40,length=100)))
{
# First, calculate true mean of truncated variable from P_theta:
Z <-pmin(rnbinom(1000000,mu=nb_mu,size=nb_disp)
,truncation_level)
truncated_truemean <- mean(Z)
# Next, generate a list of sample means and sd's
# from i.i.d. draws from the Negative Binomial distribution
#  with mean nb_mu and dispersion nb_disp:
samplemean <- rep(0,iter)
samplesd <- rep(0,iter)
for(count in 1:iter)
{
X <- pmin(rnbinom(n=SampleSize,
mu=nb_mu,size=nb_disp),truncation_level)
# Trucate values
samplemean[count] <- mean(X)
samplesd[count] <- sd(X)
```

```
}
#Enforce that our sigma' is >= d
samplesd<-pmax(samplesd,d)
# Find min alpha for P_theta
alpha_list <- abs(samplemean - truncated_truemean)
/(samplesd/sqrt(SampleSize))
a_val <- quantile(alpha_list,probs=0.95)
if(a_val > max_a_val) max_a_val <- a_val
}
}
print(max_a_val)
```

Appendix 3: R-code for Bernstein-based Confidence Interval in Section 9

One can use the following trivial R-code to compute the 0.95-quantile and the corresponding confidence interval for μ. Here the function *leftlim* maps a value for W and σ^2 and sample size n into the left-limit of the confidence interval for μ and the left-limit for the confidence interval for the true total count D of (say) violent deaths, using the a priori specified extrapolation factor $N/E|C|$ to scale it up. The function *confintdata* takes the data set $(x_1, \ldots, x_n)$ and W and maps it into the approximate confidence intervals for μ and D by using as upper bound for σ^2 the standard unbiased estimate for the variance s^2.

```
q095<-function(w,sigma2,n)
{ b<- w*log(40)/(3*sqrt(n))
b1<-2*w*log(40)/(3*sqrt(n))
b+0.5*sqrt(b1^2+8*sigma2*log(40))
}

leftlim<-function(w,sigma2,n,scaleup)
{ a<-q095(w,sigma2,n)/sqrt(n)
leftcount<-mean(x)-a
lefttot<-leftcount*scaleup
c(leftcount,lefttot)
}

confintdata<-function(w,x,scaleup)
```

```
{
n<-length(x)
sigma2<-var(x)
mu<-mean(x)
a<-q095(w,sigma2,n)/sqrt(n)
leftcount<-mu-a
rightcount<-mu+a
lefttot<-leftcount*scaleup
righttot<-rightcount*scaleup
c(leftcount,rightcount,lefttot,righttot)
}
```

An application to the 47 cluster counts in the data analysis yields

```
confintdata(40,x)
[1] 1.921892e+00 1.092917e+01 1.793368e+05 1.019829e+06

 confintdata(50,x)
[1] 1.572296e+00 1.127877e+01 1.467150e+05 1.052451e+06
```

In other words, the Bernstein-based method gives a confidence interval for the total count of violent deaths with left-limit 146,000 and right-limit $1.05 * 10^6$.

Below, we present some simple R-code for calculating the standard error of s^2. We estimate the standard error of s^2, as described in Section 9, by using that s^2 is an asymptotically linear estimator of σ^2 with influence curve $IC(X) = X^2 - EX^2 - 2EX(X - EX)$, so that the variance of s^2 can be approximated in first order with $\mathrm{VAR}(IC(X))/n$.

Function calculating estimate of standard error of s^2:

```
S2se<-function(x) { n<-length(x) barx<-mean(x) x2<-x^2
xc<-x-rep(barx,n) barx2<-mean(x2) x2c<-x2-rep(barx2,n)
ic<-x2c-2*barx*xc s2ic<-var(ic) s2<-s2ic/n sqrt(s2) }
```

Application of this function to the 47 counts in the data analysis yields the estimated standard error of s^2:

```
>S2se(x)
 [1] 20.46923
```

References

Alkhuzai, A. H., I. J. Ahmad, M. J. Hweel, T. W. Ismail, H. H. Hasan, and A. R. Y. et al. (2008, january). Violence-related mortality in iraq from 2002 to 2006. *The New England Journal of Medicine 358*(5), 484–493.

Bembom, O., J. W. Fessel, R. W. Shafer, and M. J. van der Laan (2008, March). Data-adaptive selection of the adjustment set in variable importance estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 231. http://www.bepress.com/ucbbiostat/paper231*.

Bennett, G. (1962). Probability inequalities for the sum of independent random variables. *J. Am. Statist. Ass.* (57), 33–45.

Bennett, G. (1963). On the probability of large deviations from the expectation for sums of bounded, independent random variables. *Biometrika* (50), 528–535.

Blyth, C. R. and H. A. Still (1983). Binomial confidence intervals. *Journal of the American Statistical Association 78*, 108–116.

Brownstein, C. A. and J. S. Brownstein (2008, January). Estimating excess mortality in post-invasion iraq. *The New England Journal of Medicine 358*(5), 445–447.

Burnham, G., R. Lafta, S. Doocy, and L. Roberts (2006). Mortality after the 2003 invasion of iraq: a cross-sectional cluster sample survey. *Lancet* (368), 1421–1428.

Burnham, G., R. Lafta, S. Doocy, and L. Roberts (2007, January). Mortality in iraq authors' reply. *The Lancet 369*(9556), 103–104.

Casella, G. and R. Berger (1990). *Statistical Inference*. Pacific Grove, CA: Wadsworth.

Chaudhuri, A. (2005). *Survey Sampling* (2 ed.). Boca Raton: Chapman and Hall/CRC Statistics, textbooks and monographs; v. 181.

Clopper, C. J. and E. S. Pearson (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika 26*, 404–413.

Crow, E. L. and R. S. Gardner (1959). Confidence intervals for the expectation of a poisson variable. *Biometrika 46*, 441–453.

Dudley, R. (1999). *Uniform Central Limit Theorems*. Cambridge, UK: Cambridge University Press.

Hicks, M. H.-R. (2007, January). Mortality in iraq. *The Lancet 369*(9556), 101–102.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Am. Statist. Ass.* (58), 13–30.

McCullagh, P. and J. A. Nelder (1998). *Generalized Linear Models* (2 ed.). Boca Raton, Florida: Chapman and Hall/CRC, Monographs on Statistics and Applied Probability 37.

Roberts, L., R. Lafta, R. Garfield, J. Khudhairi, and G. Burnham (2004). Mortality before and after the 2003 invasion of iraq: cluster sample survey. *Lancet* (364), 1857–1864.

Sterne, T. H. (1954). Some remarks on confidence or fiducial limits. *Biometrika 41*, 275–278.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. New York: Cambridge University Press.

von Schreeb, J., H. Rosling, and R. Garfield (2007, January). Mortality in iraq. *The Lancet 369*(9556), 101–102.

Wilcox, R. R. (2005). *Robust Estimation and Hypothesis Testing*. Burlington, MA, USA: Elsevier Academic Press.